8th Annual Institute for Genomics & Bioinformatics (IGB) Biomedical Informatics Training (BIT) Program Symposium

Donald Bren Hall Room 6011
May 19, 2010

Morning Session I – Chemoinformatics

The Significance of Chemical Similarity Scores

Ramzi Nasr, Pierre Baldi

As repositories of chemical molecules continue to expand and become more open, it becomes increasingly important to develop tools to search them efficiently and assess the statistical significance of chemical similarity scores. Here we develop a general framework for understanding, modeling, predicting, and approximating the distribution of chemical similarity scores and its extreme values in large databases. The framework can be applied to different chemical representations and similarity measures but is demonstrated here using the most common binary fingerprints with the Tanimoto similarity measure. We show that the distribution of Tanimoto scores can be approximated by the ratio of two correlated Normal distributions associated with the corresponding unions and intersections. This remains true also when the distribution of similarity scores is conditioned on the size of the query molecules in order to derive more fine-grained results and improve chemical retrieval. Z-scores, E-values, and p-values are derived to assess the significance of similarity scores. The framework also allows one to predict ROC (Receiver Operating Characteristic) curves, and to detect outliers in the form of atypical molecules. Numerous and diverse experiments carried in part with large sets of molecules from the ChemDB show remarkable agreement between theory and empirical results.

Ranking molecular orbital interactions

Matthew A Kayala, Chloe-Agathe Azencott, James Nowick, Pierre Baldi

The ability to anticipate the course of a reaction is essential to the practice of chemistry. This aptitude relies on the understanding of elementary mechanistic steps, which can be described as the interaction of filled and unfilled molecular orbitals (MOs). Here, we create a database of mechanistic steps from previous work on a rule-based expert system (ReactionExplorer) covering the vast majority of an undergraduate organic chemistry curriculum. We derive 21,000 priority ordered favorable elementary steps for 7800 distinct reactants or intermediates. All other MO interactions yield 106 million unfavorable elementary steps. To predict the course of reactions, one must recover the relative priority of these elementary steps. First, we introduce a heuristic reaction proposer based on frontier orbital theory to reduce the number of unfavorable elementary steps while maintaining high sensitivity. Then, we utilize machine learning techniques to successfully rank the proposed MO interactions. Initial cross-validated results using a neural network give a pairwise comparison accuracy of 98.9%. The quality of our database, heuristic, and ranking machinery make it an invaluable resource for the prediction of elementary reactions, and therefore of full chemical processes.

Morning Session II – Proteomics and Drug Discovery

A Molecular Dynamics Study of the AQP0-CaM Interaction

Daniel Clemens, J Alfredo Freites, Katalin Kalman, Doug Tobias, and James Hall

Aquaporin 0 (AQP0, previously MIP) is a water transport channel comprising ~60% of protein in lens cell membranes. The water permeability of AQP0 is regulated by intracellular calcium levels through calmodulin (CaM) binding. Altered sensitivity to calcium levels in response to phosphorylation of the C-terminal α-helix has also been observed. The purpose of this study is to elucidate the molecular mechanisms by which AQP0 is regulated by CaM. Specifically, we use atomistic molecular dynamic simulations to explore how various states of phosphorylation influence the interaction between the C-terminal domains of AQP0 in the native (tetrameric, membrane embedded) form with CaM .

Our simulations show that CaM can interact with the C-terminus of AQP0 stably in more than one conformation, indicating that there may be various conformational states of the AQP0-CaM complex resulting from different physiological conditions. Furthermore, our simulations of the AQP0-CaM complex confirm previous hypotheses that AQP0 is regulated by calcium through the occlusion of the transmembrane pores of the AQP0 tetramer by CaM. Additionally, our simulations give insight to the molecular mechanisms behind the altered calcium sensitivity of AQP0 in various states of C-terminal helix phosphorylation.

Annexin A2 as a model for structure-based drug design

Gabriel Ozorowski, Hartmut Luecke

Annexin A2 (AnxA2), a 36-kDa protein of the Ca^{2+}-dependent phospholipid-binding annexin family, contains F-actin binding sites in its carboxyl domain. F-actin aggregation is enhanced when AnxA2 is incubated with withaferin A (WA), a natural compound found in solanaceous plants. At the cellular level, WA has been shown to inhibit angiogenesis and metastasis. These findings make AnxA2 a potential drug target. We are employing a combination of x-ray crystallography and *in silico* docking techniques to pursue rational drug design. To determine the binding mechanism of WA to AnxA2, we attempted to cocrystallize the drug with the protein. Data from electrospray-ionization mass spectrometry has shown that the drug binds covalently to annexin A2, and this bond can be at least partially reversed with the reducing agents, such as DTT. Cocrystallization has proved to be successful in producing crystals, some of which have different unit cell dimensions than the native structure, a possible sign that a drug-protein interaction is occurring. However, the crystals only diffract to about 4Å resolution and density for the drug is not seen. Once a crystal structure is solved, we will use the information in docking simulations to screen for other potential drugs.

Virtual drug screening for the heme degradation pathway in Mycobacterium tuberculosis

Paul Rigor, Robert Morse, Celia Goulding, Pierre Baldi

Although there are several open-source and commercially available computational tools for virtual drug screening -- including Dock, Autodock and Schroedinger's Maestro; there is still a lack of a more general, tool agnostic and scalable framework that is able to leverage the advantages offered by readily available docking and molecular dynamics programs in a high-performance computing (HPC) environment. We have developed a framework built on top of an HPC pipeline and existing proteomics and chemical informatics tools -- such as ChemDB, Frag3D, SCRATCH -- to support an iterative virtual screening methodology. We've applied our approach to two biological problems and describe preliminary results: 1) putative heme degradation enzyme critical to the survival of Mycobacterium tuberculosis, and 2) a potential target involved in cancer metastasis. Moreover, future extensions to the pipeline and related tools will be discussed.

Prediction of protein antigenicity using microarray data

Christophe Magnan, Michael Zeller, Matthew Kayala, Philip Felgner, Pierre Baldi

Discovery of novel antigens is a challenging problem. Most of the existing methods are based on the sequence homology with known antigens, yet development of sequence-based predictions is limited by the small number and the high similarity of these antigens.

Here we use antigens discovered via human reactivity protein microarray experiments covering five pathogens as well as those previously noted in the literature to train a sequence-based prediction model. The overall accuracy of the resulting predictor, ANTIGENpro, estimated by multiple runs of 10-fold cross-validation is 76%. When trained from microarray data only, the same predictor correctly classifies 82% of the antigens with less than 30% similarity to any protein used to train the model. ANTIGENpro is the first sequence-based method for predicting the whole protein antigenicity from a relatively large and non-redundant set of proteins obtained by immunological microarray analysis.

Afternoon Session I – Genomics

Transcriptome-wide Quantitative Analyses of RNA Polyadenylation in Mouse Embryonic Stem Cells and Differentiated Cells using Poly(A) Site Deep Sequencing (PAS-Seq)

Peter Shepard, Eun-A Choi, Jente Lu, Lisa Flanagan, Klemens Hertel, Yongsheng Shi

Polyadenylation is an essential step in the expression of almost all protein-coding genes in eukaryotes. Additionally, a significant proportion of mammalian genes expresses multiple alternatively polyadenylated (APA) mRNA isoforms, indicating that APA is an important venue for gene regulation. Indeed, several recent genomic analyses showed that APA is highly regulated by the proliferation capacity of the cells as well as by extracellular stimuli.

We introduce a deep sequencing-based technique called PAS-Seq for transcriptome-wide quantitative analysis of RNA polyadenylation. By analyzing these poly(A) events, we detected APA in ~57% of all mouse genes in these 3 cell types, which is significantly higher than the previous estimate based on the entire mouse EST database. Transcripts of over 600 genes display significantly different APA profiles that are not dependent on splicing changes (Fisher exact test, false discovery rate 5%). For the majority of these transcripts, polyadenylation shifted toward downstream sites from ES to NPS cells (64%) and from NPS cells to neurons (73%). This is consistent with the interpretation that proliferating cells tend to use upstream poly(A) sites, i.e. shorter 3'-UTRs.

Prediction of miRNA targets in Anopheles gambiae using a word-based comparative genomics approach.

Augustine Dunn, Anthony James and Xiaohui Xie

MicroRNAs act as small, single-stranded RNAs in conjunction with protein complexes through binding to the 3'-untranslated regions (UTR) of target mRNAs to reduce the amount of target translation. They are generally ~20-22 nucleotides long and exhibit perfect Watson:Crick binding only along a seed region (6-8 nt) located at the 5'-region of the miRNA. The remainder of its interactions with the target mRNA are rich in secondary structure and can vary from target to target aggravating efforts to predict miRNA targets *in silico*. Some strategies identify seed matches, then attempt to model the binding energies of possible secondary structures of the rest of the miRNA. Others use comparative genomics to predict conserved seed regions through global alignment of orthologous 3'-UTRs.

The currently sequenced species of mosquitoes present a challenge to alignment-based comparative genomics techniques due to their long divergence times (up to 200 million years to last common ancestor) and high levels of repetitive genomic elements like transposons (~47% of the *Aedes aegypti* genome consists of transposable elements). We present here an effort to use a method based on position-independent conservation of seed regions in orthologous 3'-UTRs to predict a subset of miRNA targets in the malaria vector *Anopheles gambiae.*

Detecting Structural Variation in Natural Populations Via Paired-End Sequencing

Julie Cridland, Kevin Thornton

Structural variation has been recently associated with a number of complex traits; including several human diseases. However, little is currently known about the evolutionary dynamics of structural variants segregating within natural populations. New technology, such as Illumina's high-throughput, paired-end sequencing platform, makes it possible to study this variation by quickly and accurately detecting structural variants, even with relatively low coverage. This technique has several advantages over previous microarray experiments to detect variation, such as the ability to perform this type of experiment for any species of interest and the direct observation of sequence that is involved in structural variation.

We performed paired-end sequencing on multiple isofemale lines from population samples of two *Drosophila* species, *Drosophila melanogaster* and *Drosophila yakuba.* Many genes, such as *Cyp6g1*, which is known to be under positive selection, and *Or22a, Gr28a* and *Gr28b*, which are known to be polymorphic for copy number, were detected with this method. Hotspots of duplication were also identified throughout the genomes. We have also found evidence of natural selection acting on duplications and that duplications segregating within populations may be deleterious.

Afternoon Session II – Genomics

Genome-wide maps of candidate regulatory motif sites

Kenny Daily, Paul Rigor, Sholeh Forouzan, Yimeng Dou, Xiaohui Xie, Pierre Baldi

For any given genome, only a small fraction of the regulatory elements have been characterized, and there is great interest in applying computational techniques to systematically discover these elements. Such efforts have been hindered by the size of non-coding DNA regions and the statistical variability and complex spatial organizations of mammalian regulatory elements. A central challenge of biology is to map and understand the role of the 98% noncoding regions of the human genome. MotifMap aims to provide a comprehensive map of potential regulatory elements in genomes in an unbiased manner. Each motif found has a variety of associated scores that can aid filtering to find true, novel binding sites. One of these scores is a novel measure of conservation that takes into account not only the strength of the binding site in the species being searched, but of the strength of binding sites in related species. The methods of the MotifMap pipeline from the originally available human genome were applied to the genomes of eight other species. More species can be added in a nearly fully automated fashion, and each species is updated when novel transcription binding matrices become available through updates to existing databases or literature searches.

A Map of Known Regulatory Motifs in Fly

Jacob Biesinger, Kenny Daily, Paul Rigor, Rahul Warrior, Pierre Baldi, Xiaohui Xie

Achieving a genome-wide map of regulatory elements is a fundamental challenge in understanding eukaryotic development. As additional sequence becomes available, more refined computational methods for transcription factor binding site prediction are desired in order to take full advantage of sequence information.

To better characterize the genetic regulatory network in Drosophila melanogaster, we propose the Bayesian Branch Length Score (BBLS) as a more refined means of detecting binding sites for known transcription factors. The BBLS relaxes the assumption of a perfect sequence alignment and allows drift in the TF binding profile. Using 144 motifs representing 74 transcription factors, we predict tens of thousands of binding sites in the fly genome. A genome-wide ChIP-chip benchmark shows increased sensitivity and specificity. Further, our method predicts more sites with higher confidence than previous methods. We use these predictions to investigate the Dpp signaling pathway, validating several new, predicted binding sites for the SMM transcription factor. We have made our predictions available for download, browsing and visualization through the MotifMap web server at http://motifmap.ics.uci.edu. Predictions can be integrated with either the UCSC browser or the FlyBase genome browser through our web service.

Co-factor of LIM Transcriptional Regulation in Mammary Gland Development

Michael Salmans, Padhraic Smyth, Bogi Andersen

Co-factor of LIM (Clim) proteins recruit LIM domain transcription factors to mediate their interaction with DNA targets. Expression of the Clim2 isoform in the mammary gland is essential as evidenced by transgenic mice expressing a dominant-negative Clim molecule (DN-Clim) under the epithelial-specific control of the K14 promoter. Mammary glands in DN-Clim mice exhibit decreased branching morphogenesis and terminal end bud (TEB) size, delayed ductal elongation, and depletion of the stem cell population. Preliminary gene expression profiling of whole mammary glands from eight week old wild type and DN-Clim mice suggests Clims regulate pathways required for stem cell maintenance, TEB formation, and branching morphogenesis. To further investigate the transcriptional role of Clims in the mouse mammary gland we have isolated TEB and ductal tissue by laser capture microscopy for a time course gene expression profiling of several key stages during pubertal development. Computational analyses will reveal the gene networks that operate during normal mammary gland development throughout puberty and those specifically under transcriptional regulation by Clims. Chromosome immunoprecipitation followed by DNA sequence analysis (ChIP-Seq) will reveal the sequence motifs associated with Clim2 binding and provide insights into the various transcription factors under control of Clim2.

Mechanisms of Transcriptional Regulation by Cohesin in Cornelia de Lange Syndrome

Aniello Infante, Daniel Newkirk, Kyoko Yokomori, Xiaohui Xie

Cohesin, together with NIPBL, plays an integral role in sister chromatid cohesion during mitosis, but it has also been implicated in gene regulation; however, the mode of action is not known. Mutations to the SMC components of cohesin lead to a number of developmental disorders, and mutations in NIPBL can lead to Cornilia de Lange syndrome. In this study we use ChIP-seq with rad21 (a cohesin subunit) and NIPBL (a cohesin loader) in wildtype mice and a heterozygous NIPBL mutant, in MEF cell lines and brain tissue. We identify genome-wide binding sites for both cohesin and NIPBL, and explore their properties. As expected, the mutant strain loses a large number of cohesin binding sites. Surprisingly, it also gained novel sites not seen in the wildtype. We also perform a motif search over the cohesin binding sites, and discover a CTCF motif; a genome-wide analysis of CTCF and cohesin binding shows significant overlap, as has been previously shown. We compare the differences in binding sites to gene expression levels in both wildtype and mutant tissues, and show a significant enrichment between promoter binding by cohesin and expression levels.

Afternoon Session III – Systems Biology

Sigmoid: Spatial and Compartmental Modeling for Pathway Bioinformatics and Systems Biology

Ben Compani, Trent Su, Ivan Chang, Pierre Baldi, Eric Mjolsness

Progress in systems biology critically depends on developing scalable informatics tools to model and visualize complex biological systems, and to flexibly store information about these systems and their models.

Here we describe *Sigmoid*, a generative, scalable software infrastructure for pathway bioinformatics and systems biology. Sigmoid uses the web services framework to create a distributed modeling system. This flexible framework offers powerful modularity that, in conjunction with the generative nature of the

Sigmoid coding cycle, offers a significantly reduced development time for integration of new components. Sigmoid capitalizes on the robust mathematical software tools and problem solving environment that Mathematica offers, along with the Xcellerator/kMech/Cellzilla packages designed to facilitate biological modeling via automated equation generation. To address the need to develop multi-cellular models, recent progress adds classes and data structures for spatial and multi-compartmental modeling including support for Cellzilla. The synthesis of these features yields a flexible scalable architecture that not only allows for manageable adoption of new system components, but may open the ability to play within yet larger bioinformatics frameworks.

Graph-Constrained Correlation Dynamics: Approximation of dynamical systems

Todd Johnson, Eric Mjolsness

We present a general, goal-oriented approximation framework which for the optimization of approximations to a relationship between a set of fine-scale variables x and a set of observables z, in the form of a relationship between a set y of course-scale variables and the same set of observables z.

Consider the particular case where the relationship between the fine-scale variables and the observables evolves continuously in time, but calculating the conditional probability for arbitrary time points is difficult, such as a system with a large state space governed by a chemical master equation. In such a situation, we provide a method of finding an optimal approximation within a family of distributions described by a 1) a predeterimined Markov Random Field and 2) a set
basis functions for the derivatives of the parameters of the MRF. This approximation is optimal under a time-averaged Kullback-Leibler divergance, and allows the modeling of specific covariances of interest along a tractable time-evolution of the probability distribution.

We motivate this work with a model of CaMKII binding in synapses, for which we present a dynamical grammar model and a corresponding Markov Random Field model optimized by GCCD.

Incorporating Existing Network Information into Gene Network Inference

Scott Christley, Qing Nie, Xiaohui Xie

Traditionally, gene regulatory networks (GRN) have been experimentally constructed using a combination of mapping of transcription factor (TF) binding sites, physical binding of TFs to these sites, and demonstration of the importance of both to gene expression output. This “pipeline” is both labor and time intensive, requiring tens of man-years of effort to build an extensive network. Another approach, utilizing computational methods operating on gene expression data, can produce a GRN with the advantage of rapidly generating a global perspective. One successful method for inference of a GRN is based upon ordinary differential equations (ODE) describing gene regulation as a function of other genes. Furthermore, new data is being produced such as ChIP-chip and ChIP-seq experiments that measure the physical interactions between transcription factors and the genes they regulate, so incorporation of this data as a priori network information should enhance the inference procedure. We have developed a general optimization framework based upon the ODE method that incorporates existing network information in combination with regularization parameters that encourages network sparsity. Our framework can utilize a variety of gene expression experiments such as perturbation, null mutant and heterozygous knockdown. We demonstrate our method on simulated network data and experimental data for mouse embryonic stem cells.